DOCUMENT RESUME

ED 237 934 CS 007 418

AUTHOR	Vinsonhaler, John F.; And Others
TITLE	Improving Diagnostic Reliability in Reading through Training. Research Series No. 126.
INSTITUTION	Michigan State Univ., East Lansing. Inst. for Research on Teaching.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
REPORT NO	IRT-RS-126
PUB DATE	Jul 83
CONTRACT	400-81-0014
NOTE	42p.
AVAILABLE FROM	Institute for Research on Teaching, College of Education, Michigan State University, 252 Erickson Hall, East Lansing, MI 48824 ($4.00).
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Clinical Diagnosis; Elementary Secondary Education; Individualized Instruction; *Interrater Reliability; Models; Reading Centers; *Reading Consultants; *Reading Diagnosis; Reading Instruction; *Reading Teachers; *Remedial Reading; *Training Methods

ABSTRACT

While diagnosis is generally considered a vital element in reading clinicians' expertise, research has revealed that even degreed, experienced reading clinicians display little personal consistency or agreement with one another when diagnosing simulated cases of reading difficulty. Three studies were conducted to determine if systematizing the diagnostic process by providing a process model, diagnostic decision aids, and sufficient practice with feedback would result in more reliable diagnoses. Subjects were (1) master's degree students in reading who had some prior course work in diagnosis, (2) master's degree candidates with prior teaching experience and coursework in reading, and (3) experienced classroom teachers with little or no formal training in reading or reading diagnosis. The results indicated that the training was successful, both with degreed reading clinicians and with teachers who had no previous work in reading diagnosis. (Appendixes contain the cue inventory for one simulated case, a portion of a diagnostic decision aid, a portion of a diagnostic checklist, and tables of data from the studies.) (Author/FL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED237934

Research Series No. 126

IMPROVING DIAGNOSTIC RELIABILITY IN READING THROUGH TRAINING

John F. Vinsonhaler, Annette B. Weinshank,
Ruth M. Polin, and Christian C. Wagner

Published By

The Institute for Research on Teaching
252 Erickson Hall
Michigan State University
East Lansing, Michigan 48824

July 1983

This work is sponsored in part by the Institute for Research on Teaching, College of Education, Michigan State University. The Institute for Research on Teaching is funded primarily by the Program for Teaching and Instruction of the National Institute of Education, United States Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position, policy, or endorsement of the National Institute of Education. (Contract No. 400-81-0014)

CS 007413

Institute for Research on Teaching

The **Institute for Research on Teaching** was founded at Michigan State University in 1976 by the National Institute of Education. Following a nationwide competition in 1981, the NIE awarded a second contract to the IRT, extending work through 1984. Funding is also received from other agencies and foundations for individual research projects.

The IRT conducts major research projects aimed at improving classroom teaching, including studies of classroom management strategies, student socialization, the diagnosis and remediation of reading difficulties, and teacher education. IRT researchers are also examining the teaching of specific school subjects such as reading, writing, general mathematics, and science, and are seeking to understand how factors outside the classroom affect teacher decision making.

Researchers from such diverse disciplines as educational psychology, anthropology, sociology, and philosophy cooperate in conducting IRT research. They join forces with public school teachers, who work at the IRT as half-time collaborators in research, helping to design and plan studies, collect data, analyze and interpret results, and disseminate findings.

The IRT publishes research reports, occasional papers, conference proceedings, a newsletter for practitioners, and lists and catalogs of IRT publications. For more information, to receive a list or catalog, and/or to be placed on the IRT mailing list to receive the newsletter, please write to the IRT Editor, Institute for Research on Teaching, 252 Erickson Hall, Michigan State University, East Lansing, Michigan 48824-1034.

Co-Directors: Jere E. Brophy and Andrew C. Porter

Associate Directors: Judith E. Lanier and Richard S. Prawat

Editorial Staff
Editor: Janet Eaton
Assistant Editor: Patricia Nischan

Abstract

Diagnosis is generally considered a vital element in the expertise of reading clinicians. Yet our previous research revealed that even degreed, experienced reading clinicians displayed very low agreement with themselves and with one another when diagnosing simulated cases of reading difficulty. This paper reports the results of three studies designed to see if systematizing the diagnostic process by providing (1) a process model, (2) diagnostic decision aids, and (3) sufficient practice with feedback would result in reliable diagnoses. The results indicate that the training (which can easily be incorporated into typical courses in reading diagnosis) was successful, both with degreed reading clinicians and with teachers who had no previous course work in reading diagnosis.

IMPROVING DIAGNOSTIC RELIABILITY IN READING THROUGH TRAINING

John F. Vinsonhaler, Annette B. Weinshank,
Ruth M. Polin, & Christian C. Wagner[1]

A major reason for studying diagnosis of reading difficulties is the importance accorded it by nearly all authorities in reading. Diagnosis as the basis for remediation is an important principle in the literature and in practice (Carter & McGinnis, 1970; Ekwall, 1976; Otto, McMenemy & Smith, 1973; Rabinovitch, 1965; Smith, 1969; Smith, Carter & Dapper, 1970; Spache & Spache, 1973).

At least three major orientations toward diagnostic content can be found in the literature. Advocates of one approach establish general reading levels compared to reading potential (Guszak, 1972; Spache, 1976). Advocates of a second view emphasize performance on a set of reading skills. Advocates of a third approach use diagnosis as determination of causality, that is, understanding the underlying factors that have caused reading problems. Such an understanding supposedly enables the clinician to prescribe the most appropriate steps for remediation (Carter & McGinnis, 1970; Harris, 1972; Harris, 1977; Monroe, 1968; Natchez, 1968; Strang, 1964).

Regardless of the content of reading diagnosis, nearly all authors agree that the diagnosis should form the basis for remediation. However, with few exceptions (Bateman, 1971; Spache, 1969), authors have not dealt with the

---

[1]John F. Vinsonhaler and Annette B. Weinshank coordinate the Outcomes in Reading Project. Vinsonhaler is a professor in the Counseling, Educational Psychology and Special Education Department at MSU. Weinshank is a teacher collaborator with the IRT. Ruth M. Polin is data processing coordinator for the project. Christian C. Wagner is a consultant to the project and is now at the College of Engineering at Oakland University. The authors gratefully acknowledge the helpful comments and leadership of IRT Co-Director Jere Brophy.

effect of unreliable diagnosis on development of knowledge about treatment outcomes.

Consider, for example, a study in which two remediations for a given diagnostic category are being evaluated. If reading diagnosticians demonstrate low reliability, identifying which type of problem a student has is essentially a random choice. Assume that one of these remediations is effective. This effective treatment will improve performance, but only for those students who happen to have been diagnosed correctly. Overall, to the degree that the diagnoses are unreliable, the efficacy of a differentially effective treatment will be systematically underestimated. Furthermore, reliability of diagnosis does not necessarily inform validity (one can be reliably wrong). Reliability does, however, permit the correct estimation of remedial effectiveness (Cohen, Rubin, Neyman, Dantzig, Baer, & Siegelaub, 1964).

Empirical Studies of Diagnosis

There are conflicting reports in the medical literature on the agreement among individual physicians on medical judgments. Several studies indicate substantial agreement among physicians; others show marked disagreement (Cochrane & Garland, 1952; Fletcher, 1952; Garland, 1959; Paton, 1957; Yerushalmy, 1955, 1969). For example, Lerner and Schuyler (1973) suggest that groups of clinicians, working together, can produce diagnostic statements that are mutually agreed upon. Educational clinicians, working alone, however, yield less promising results.

In a series of observational studies, we analyzed the written diagnoses and remedial plans of reading specialists and special-education clinicians to determine commonality (group agreement) and individual agreement about simulated cases (see Vinsonhaler, Weinshank, Wagner, & Polin, 1983). The

initial study revealed very low agreement among specialists and in individual diagnoses. This finding was startling, considering that the subjects were experienced, highly regarded reading clinicians. We performed a series of five additional observational studies to see if these unexpected findings could be replicated and generalized from. We drew new samples from additional populations, including other reading specialists, classroom teachers, and learning disabilities clinicians. In addition, we developed and used new simulated cases and case formats. Potential errors that could result from the translation of written diagnoses to standardized categories were eliminated through use of a standardized diagnostic checklist. Finally, we investigated the reliability of diagnostic categories that were linked to suggested remediations and of the remediations themselves (Weinshank, 1982). Individual diagnostic and remedial reliability remained very low across all the studies (i.e., clinicians very frequently disagreed about what the problems were and how to remediate them). Mean interclinician reliability averaged 0.03 (Phi) and 0.08 (Porter). Mean intraclinician reliability averaged 0.21 (Phi) and 0.20 (Porter). The initial findings on commonality were also confirmed. Mean commonality across the studies was only fractionally higher than the minimum possible value.

These studies show that, as a group, education professionals, including reading specialists, produce diagnoses, the content of which shows in aggregate some signs of conforming to the recommendations found in the literature. That is, diagnoses usually included statements about reading potential, strengths and weaknesses in skills, and suspected causal factors (hearing, vision, and attitude).

Individually, however, the diagnoses show significant deviations from the recommendations in the literature. First, they include a large number of one-time-only statements of questionable relevance to remediation. Second, they systematically fail to mention the reading skills of greatest import to remediation. Third, even when important skills are mentioned, these statements are not reliably linked with treatment prescriptions (Weinshank, 1982).

One explanation for this unreliability might be that these studies used simulated cases in an experimental environment. However, the use of actual children in a natural setting might further decrease agreement, since a child's performance would be expected to change, thereby introducing unreliability in the data base.

The differential effects of using real and simulated cases has been studied in medicine. No differences were found when the diagnoses were compared for (1) people with real medical problems and (2) people coached to simulate the same medical problems (human simulation). Further, in studies comparing human simulation of medical problems with simulated cases whose format was similar to those used in our studies, differences were found in procedure, but not in the final diagnoses (Norman & Tugwell, 1981).

We favor a second explanation for the low diagnostic agreement found in our studies: Reading specialists receive inadequate training. A comparison of training programs in medicine and reading is instructive here. Medical training is based on (1) an organized body of empirically based knowledge that relates specific remedies to specific problems (Copp, 1976; Johnson, 1975; King, 1976; Puck, 1976; Roos, 1975); (2) systematic techniques governing the collection of cues (DeDombal, Leaper, Horrocks, Staniland, & McCann, 1974; Elstein, Shulman, & Sprafka, 1978; Prior, Silberstein, & Stang, 1981); and (3) perhaps most importantly, the supervised diagnosis, treatment, and follow-up

of thousands of cases (Shapiro & Lowenstein, 1979; Simpson, 1972). By contrast, training in reading diagnosis and remediation is based on (1) non-empirically verified theoretical concepts, (2) idiosyncratic cue collection techniques, and (3) supervised diagnosis, remediation, and follow-up on few cases.

Diagnostic Training Hypothesis

Here we report the results of three studies investigating the diagnostic training hypothesis that improved clincial training can increase diagnostic reliability.

A Theory of Clinical Problem Solving

The problem-solving behavior of clinicians in medical and other professions has led to a theory of how diagnostic and treatment decisions should be made (DeGowin & DeGowin, 1976) and to observation of how they actually seem to be made (Bordage, 1982; Elstein, Shulman, & Sprafka, 1978). According to the theory, there are two participants in the clinical problem-solving setting. The first is any complex system (whether an interaction between a case and a clinician or an individual and a clinician) referred to as a case. The proper functioning of the case is inferred from its performance on certain critical variables.

The second participant is a problem solver, the clinician. The clinician maintains cases and tries to improve the case (a human's) performance. The interaction between clinician and case is usually initiated by a problem with case performance (DeGowin & DeGowin, 1976). The actions taken by the clinician have been organized around the terms "diagnosis" and "treatment" and are all logically based upon the clinician's model of process (i.e., how critical performances and causal factors are related).

The principal explanatory device in the empirical theory is clinical memory. Memory consists of (1) associations between cues (case information) and potential problems and (2) associations between problems and treatments. Decisions are driven by hypothesis testing (i.e., the generation of a set of likely problems and the collection of cues to rule in or rule out the hypothesized problems). The principal means of validating this theory is artificial intelligence (e.g., computer generated diagnoses) and computer-simulation experiments, which predict the problem-solving behavior of real clinicians. The behavior predicted is the making of diagnostic judgments, given case information. Such studies in reading (Gil, Wagner & Vinsonhaler, 1978; Wagner, 1982) show that the theory predicts the reliable portions of reading clinicians' behavior.

Because the heart of the theory is clinical memory and clinical memory is dependent on a model of process, it follows that a model of the reading process must form the basis for the design of a training program in reading.

A Model of Reading

The model chosen to guide the training studies described here is the Model of Reading and Learning to Read (MORAL) first developed by George Sherman of Michigan State University and subsequently expanded and adapted for these studies (Cureton, Stewart, & Patriarca, 1980; Weinshank, Cureton, & Blatt, 1980). The model describes a series of critical performances that a skilled reader must demonstrate, together with the concurrent cognitive skills, personal and environmental factors, learning history, and learning skills that would enable and sustain the critical performances.

In this training model the reader (1) receives input from the environment; (2) processes this input in conjunction with his/her own memory of past

events; and (3) produces an output that affects its memory, the environment, or both. A particular reader, for example, attempting a particular reading task, receives as input the requirements of the task. This input, together with past knowledge of reading and language, are processed in some way, and outputs are produced. Some effects are not observable (e.g., changes in memory) and some are (e.g., performance on the reading task as measured in some way).

In our training studies, we found seven reading and language performances critical to effective reading. To the degree that these performances are inadequate, mastery of some reading tasks may be impeded.

1. instant word recognition performance, defined as the ability to recognize a certain set of words instantly
2. decoded word recognition, defined as the ability to recognize a set of words using various association strategies (e.g., sound-symbol association)
3. vocabulary, defined as the ability to give word meanings
4. oral reading, defined as the ability to read text aloud with appropriate phrasing, fluency and intonation
5. silent reading comprehension, defined as the ability to answer specific questions on text read silently
6. listening comprehension, defined as the ability to answer specific questions on text read aloud by someone else
7. attention/motivation, defined as the ability to activate and maintain concentration on the task at hand

The MORAL goes further than specification of the critical performances. For each critical performance this model specifies the associated causal factors (i.e., the child and environmental factors that affect his or her performance). For example, if the child has poor instant word recognition, the MORAL suggests investigating probable causal factors such as poor visual discrimination, insufficient reading practice, and so on.

Requirements of Effective Diagnostic Training

Three features characterized the effective diagnostic training used in these studies. First, instruction must provide training on a *model of the reading process* to serve as the foundation for the organization of clinical memory. Second, instruction should include training with *decision aids* to insure systematic data collection and diagnostic decision making. Finally, *practice with feedback* is necessary to consolidate clinical memory and strategy.

The MORAL provided a training process for our studies. Clinicians were taught to use the MORAL to (1) identify the most important reading performances and (2) infer significant underlying causes of those performances. For the studies reported here, we developed decision and training aids by examining the most likely causes for all of the critical reading performances. From these we devised lists of inferences (see Table 1). Two major categories of aids were created: diagnostic/remedial forms for use during diagnostic decision making and diagnostic checklists for translating written diagnoses into a common vocabulary. We also provided extended practice with feedback on decisions. A senior clinician, operating in accordance with the model, evaluated study participants' diagnoses of several cases.

Elaboration and refinement of these training elements occurred over the course of the three studies reported here.

The Initial Training Study (1977)

The purpose of the initial training study was to investigate the effects of non-model based training on the participants' agreement with a criterial diagnosis. Specifically investigated was the impact of non-model based

Table 1

Critical Reading Performances and Examples of Causal Factors

Critical Reading Performance	Examples of Causal Factors
Instant Word Recognition	Visual discrimination of words; Visual memory of words; Decoded word recognition skills
Decoded Word Recognition	Auditory memory and discrimination; Segmentation/blending; Use of context
Word Comprehension	Word knowledge; Verbal concepts
Reading Comprehension	Instant word recognition; Decoded word recognition; Word comprehension; Processing strategies
Listening Comprehension	Text comprehension frames and strategies
Oral Reading	Instant word recognition; Decoded word recognition; Word comprehension
Attention/Motivation	Amount and condition of effective practice; Attention of the learner; Relevance (transferability) of practice task; Learner's correct perception of the task; Corrective feedback

decision aids and practice. The participants were master's degree students[2] in reading who had already taken some prior course work. Clinical training

[2]To avoid confusion, participants in the study will be referred to as participants or students. A student diagnosed to have a reading problem will be referred to as a child.

consisted of 30 hours of instruction in a five-week class format. There were three groups for which the treatments differed. One group used real cases, the second used simulated cases, and the third used simulated cases with decision aids (diagnostic flow charts).

Materials

The stimulus materials used for testing and for training were four different simulated cases of reading difficulty which had been used in the observational studies of reading specialists described above (Vinsonhaler et al., 1983). Each case was based on data from a child who had attended the Michigan State University Reading Clinic. The four simulated cases were representative of reading problems commonly encountered in public schools. Grade levels in the cases ranged from third to seventh.

A variety of problems were covered, including: depressed sight vocabulary, inadequate oral reading fluency, problems with application of decoding skills and with decoding of multisyllabic words, high frequency hearing loss, and comprehension problems involving the demands of content-related materials.

All cases included an audiotaped interview with the child and a brief statement of the reason for referral to the clinic (typically, below grade placement performance in reading-related subjects). The rest of each simulated case consisted of all the information (cues) that had been collected during testing sessions with that child. At the time the cases were developed, the Reading Clinic was choosing from among a variety of formal and informal measures to collect information about the children's home, school, and physical background; cognitive ability; academic achievement; and individual reading performance. The items of information collected for each case (completed forms, test scores, test booklets, examiner's comments and audiotapes)

were stored in a portable file box. A cue inventory listing all the information available was provided for each case. The cue inventory for a simulated case (Case 4: Dan) is shown in Appendix A.

Each simulated case had an equivalent form--a superficially disguised replicate of the original prepared by changing the child's name, using alternate forms of tests, and so on (Lee & Weinshank, 1978). Thus, there were four original cases and four replicates.

Design

The design involved pre- and posttesting on a randomly assigned simulated case. There was no control group because diagnostic agreement was known to be stable at a very low level. The dependent variable was agreement with a diagnosis prepared by three senior clinicians working as a group.

The criterion diagnosis was a set of weights assigned to each stated diagnostic category. Higher weights were assigned to categories judged important by the group. The child's score was the sum of the weights for the categories mentioned in the child's diagnosis divided by the sum of the weights on the categories in the criterial diagnosis. For example, suppose the criterial diagnosis included sight work (with a weight of 1.0) and poor oral reading (with a weight of 0.5). If a child's diagnosis included sight words and poor comprehension, the child's score would be 1.0 divided by the sum of the clinicians' weights (1.0 plus 0.5).

Results

Training substantially improved diagnostic agreement. Mean pretest agreement with the criterial diagnosis was 0.16, while mean posttest agreement was 0.46. There were no marked differences among the treatment groups.

Thus, this study confirmed the training benefits of systematizing information collection and using diagnostic decision aids to reach diagnostic judgments. In addition, the study served as a first approximation for the training and measurement methods used in subsequent studies.

The Second Training Study (1979)

The second study attempted to determine if further improvement in agreement (beyond that produced by the decision aids) would result from model-based training. Diagnostic agreement was measured by correlations between diagnoses of study participants for the same case, rather than agreement with a target diagnosis. We instructed participants in how to use a model of reading with four (instead of seven) critical-reading performance criteria. Decision aids and practice with instructor feedback were both based on the Model of Reading and Learning to Read.

Methods

Twenty-eight experienced teachers (master's degree candidates with prior coursework in reading) received 30 hours of training in a five-week graduate course on reading diagnosis.

Instruction

The participants received instruction based on the Model of Reading and Learning to Read (MORAL). They practiced what they had learned on four simulated cases of reading difficulty, made diagnostic decisions about the cases, wrote them up, and received the instructor's comments about their written diagnoses. These training cases were different from the ones they diagnosed in pre- and posttests. We gave the student-participants a decision aid consisting of diagnostic/prescriptive summary forms to guide their interactions with the simulated cases.

On the first day of class, we randomly assigned students to one of four groups (seven students in each group (N=28)). The four groups received identical training (e.g., the MORAL and practice on simulated cases), but were tested with different simulated cases. (These four simulated cases and their replicates were the same as those used in the initial training study.)

We conducted classroom instruction in three-hour blocks, twice weekly, for five weeks. The course topics and their order of presentation were governed by the MORAL. A handout containing the MORAL in matrix form was distributed on the first day. We gave various demonstrations during and/or following lectures. These included the administration, scoring, and interpretation of measures used to assess the critical performances of reading. A simulated case, like those used for the pretest, served as the basis for many of the demonstrations.

Students were required to work on one practice case each week (four in all). All practice cases were computer-based, but the format of the case information was the same as in the manually based cases used for pre- and posttesting. Students completed data-base and causality checklists for each case and then received feedback and had the opportunity to discuss the case during formal class time. Although the instructor gave feedback to the total class, the instructor did not examine the diagnoses made by individual students. Hence, there was no assurance that the student and instructor examined all critical performances and likely causal factors in every case.

Testing

The students received complete directions for diagnosing the simulated cases on their pre- and posttests. They read and simultaneously listened to recorded instructions about how to use a simulated case. To check their

understanding, they were to request information from a practice case different from the one they would subsequently diagnose. After instructions, the students received initial contact information about the case to be diagnosed; this included a short summary about the child's reading performance (e.g., "The child is 10 years old and reading at third-grade level."). The students then had 45 minutes to collect as many cues (items of information) about the case as they wished. The instructor asked them to list on a record form, in order of collection, all cues they selected.

After 45 minutes, they wrote their diagnoses using the categories on a diagnostic/prescriptive summary form. This form channelled their thinking and diagnostic write-up toward four of the critical performances (instant word recognition, decoded word recognition, oral reading, silent reading comprehension) and their causal factors.

Then, students were asked to match their written diagnoses with diagnostic categories listed on two different checklists. This procedure standardized the student-participants' vocabulary for comparison and data analysis. Requiring them to write their diagnoses before completing the checklists was suggested from results of prior work showing that participants given the checklist immediately tended to check off all items whether or not they characterized the case.

The MORAL data-base checklist listed 49 statements about a child's reading status. The students were to mark only those categories they had mentioned in their diagnostic write-up. For example, students who stated that a child lacked visual memory were to check Category 7, "inadequate visual memory of word forms."

The same procedures applied for the second checklist, the MORAL causality checklist that included 25 statements indicating various causes of poor

results on the four critical reading performances. For example, students who mentioned that the child had inadequate instant word recognition because s/he lacked practice should have checked the first statement, "Inadequate instant word recognition is partially caused by insufficient independent reading practice." on the causality checklist.

Identical procedures took place in the posttest session on the last class day. Students received the same case they had diagnosed in their pretest. For a complete description of procedures and materials used, see Gil, Polin, Vinsonhaler & VanRoekel (1980).

Data Analysis and Results

Data analysis focused on (1) the extent of group agreement (commonality) and (2) the percentage of diagnostic agreement among students (interclinician agreement). Agreement statistics were calculated separately for the data-base and causality checklists, and means are reported for all participants diagnosing the same case. Most participants marked almost all categories on the checklists, despite instructions to mark only those items they wrote in their own diagnoses. Therefore, all categories in the checklist that had not appeared in the students' initial diagnostic write-ups were discarded before analysis of diagnostic agreement. The reliability of this verification procedure was checked by repeating it on a random sample of the diagnostic checklists. In 85% of the decisions to discard checklist items, both coders agreed. Data analyses then were run on all checklist categories that coders verified had been included in the students' written diagnoses.

Diagnostic Commonality

Commonality results for the pretest were higher than in the observational studies (0.28 for data-base checklist and 0.26 for the causality checklist

versus 0.20 for the observational studies). This improved, group agreement probably resulted from the use of decision aids based on a model of reading, in this case, the MORAL. In addition, mean commonality increased between the pre- and posttests (from 0.28 to 0.36 for the data-base checklist and from 0.26 to 0.44 for the causality checklist). This change reflects improved group agreement resulting from model-based training with feedback. As in the previous observational studies, the commonality results again show the pervasiveness of the seven critical performances as commonly used diagnostic categories.

Interclinician Diagnostic Agreement

The mean inter-clinician correlations in Table 2 show that individual diagnostic agreement on the pretest was higher in the training study than in the observational studies. For example, for the data-based checklist, which includes judgments on the four critical reading performances, the mean initial agreements were 0.26 (Phi)[3] and 0.17 (Porter) notably higher than the 0.03 (Phi) and 0.08 (Porter) obtained for the total diagnosis in the observational studies. The higher, mean initial agreement on the causality checklist is probably due to the use of the model-based decision aids; all other conditions were identical to those common to the observational studies.

Students' diagnostic agreement improved from pre- to posttests on both checklists. Thus, model-based clinical training with feedback was effective in improving individual diagnostic agreement beyond that produced by the decision aids. Finally, the data show that a greater improvement was obtained on

---

[3]Explanations of Phi Correlation and Porter Statistic are found in Appendix B.

the causality checklist than on the data-base checklist. Students began the course with higher Phi correlations on the data-base checklist and improved less on the data-base checklist than they did on the causality checklist.

Third Training Study (1980)

The purpose of the final training study was to evaluate the improvement in diagnostic agreement that would result when the model-based training was more tightly controlled. The model used, the MORAL, included all seven of the critical reading performances. Classroom instruction in the model was based on a text developed expressly for training (Weinshank et al., 1980). The model-based decision aids were redesigned such that students were forced to (1) make a yes or no decision on the status of each critical reading performance, (2) support that decision with case data, and (3) list probable causes underlying performance (see Appendix C). Model-based practice was given with feedback specific to each student.

Methods

The 15 participants, experienced classroom teachers with little or no formal training in reading or reading diagnosis, were chosen so that we might determine the effectiveness of this type of training with non-specialists.

The student-participants were divided into three training groups, each with a different preceptor (i.e., an experienced clinician who diagnoses and remediates according to a model of process and provides feedback on student decision making for specific cases). The three groups were instructed for 30 hours and given 10 hours of extra practice time in the use of (1) the MORAL; (2) simulated and/or real cases with instructor feedback; and (3) decision aids that guided the interaction of simulated case users. Progress was monitored by means of pre-, mid-, and posttests on a simulated case, and an

Table 2
Inter-Clinician Correlations

	Data Base Checklist		Causality Checklist	
Case	Pretest	Posttest	Pretest	Posttest
1				
Phi	.32(.13)[a]	.39(.13)	.16(.27)	.38(.18)
Porter	.22(.09)	.28(.11)	.17(.16)	.39(.11)
2				
Phi	.23(.18)	.47(.12)	.14(.25)	.37(.18)
Porter	.15(.12)	.34(.10)	.13(.16)	.33(.14)
3				
Phi	.15(.17)	.31(.16)	.19(.23)	.37(.24)
Porter	.09(.10)	.23(.13)	.16(.12)	.34(.18)
4				
Phi	.33(.15)	.36(.12)	.14(.26)	.39(.20)
Porter	.22(.09)	.26(.10)	.14(.18)	.37(.16)
Grand mean				
Phi	.26(.08)	.38(.07)	.16(.02)	.38(.01)
Porter	.17(.06)	.28(.05)	.15(.02)	.36(.03)

[a]Standard deviations appear in parentheses.

additional posttest (transfer test) on a case not previously diagnosed. Five simulated cases were used; one student from each preceptor training group was tested on each case.

The materials used in this study included the same set of four simulated cases and their replicates used in the previous training studies. In addition, a new simulated case and replicate were developed to provide an example of a reading comprehension problem in an older child.

For two of the groups, the formal classroom instruction in reading diagnosis was conducted in weekly three-hour blocks with additional time spent outside the class diagnosing computer-based simulated cases (as opposed to the

manually-based ones used for the test sessions). After examining a simulated case, students filled out the decision-aid diagnosis sheets. Then they translated their diagnoses to a standardized checklist, indicating whether the case showed adequacies or inadequacies in the seven critical reading performances and their causal factors as postulated by the MORAL. Students in the remaining group, who used real, not simulated cases did not use the checklist. Instead, their preceptor analyzed the real cases diagnosed by each student in class.

Testing

The testing procedure replicated that used in the second training study except that

1. there were five simulated cases rather than four,
2. there were four testing sessions (pre-, mid-, posttest, and transfer of training) rather than just pre- and posttests; and,
3. a revised, model-based, diagnostic decision aid (discussed above) and checklist were developed using the MORAL (with its seven critical reading performances rather than four).

A portion of the decision aid is shown in Appendix C.

The decision aid forces the individual to

1. make a judgment about the adequacy or inadequacy of each critical reading performance,
2. indicate the case information used in making the decision,
3. list likely causal factors underlying performance, and
4. suggest remedial strategies.

This decision aid was based on the problem-oriented medical record developed by Weed (1976).

The MORAL required participants to say whether the child in each case performed adequately or inadequately on each category of critical reading

performance. Subsets of diagnostic categories under each critical-reading-performance category included related causal factors. Under each critical performance, an "other" category accommodated those diagnostic statements by students that could not be translated into existing categories. In addition, some causal factors related to learning were listed separately at the end of the checklist. A portion of that checklist is shown in the Appendix D.

The pretest was administered prior to any group meetings. Identical procedures were followed for the midtest (approximately five weeks later) and the posttest (at the end of 10 weeks). On each test, students diagnosed the same case; thus a progress profile was established. A week after the first posttest, a second posttest (the transfer test) was given in which participants diagnosed a different simulated case, one they had never seen. (For a complete description of procedures and materials used, see Polin, 1981.)

Results

Observations of Instructional Activities

Activities during all sessions with the three preceptors were recorded continuously, with times noted at approximately 10-minute intervals. The recorded observations were coded into three descriptive categories: (1) type of interaction, (2) topics covered, and (3) sources of topics. (Study participants will be referred to as students since they were involved in the studies as students in classes.)

Table 3 is an excerpt of an observation protocol and its translation to the coding sheet.

Table 3
Observation Excerpt

Interaction		Coding of Interaction	
Preceptor:	"What is tested if I give a child sentences to put in order?"	#2:	Preceptor questioning, student answering
Preceptor:	"Follow an analytical course. Use as few tests as possible. If instant word recognition is a problem, go to DOLCH or some such instant word-recognition test."	#1: #8: #15:	Preceptor lecturing, students listening Instant word recognition Cue collection

Table 4 shows the types of interactions preceptors and students engaged in during the 10 training sessions and what proportion of 10-minute segments from each of 10 sessions were spent in each type of interaction.

In addition, the table shows the proportion of 10-minute blocks in which each topic was observed. As can be seen, a great deal of the time was spent discussing critical reading performances.

In summary, preceptor training in this study is characterized by lecturing and question answering on a common set of topics consisting mainly of the critical reading performances. Preceptors differed in the sources they preferred to use for discussing the topics. One relies heavily on personal experiences, standardized tests, and real cases as springboards for discussion; another prefers more formal sources: written materials and simulated cases.

Individual Agreement: Students with Students Versus Students with Their Preceptors

On the basis of the earlier training studies, we expected that agreement among students on the pretest would be higher than that obtained in the

Table 4

Frequency of Occurrence of Various Interaction Types, Topics, and Topic Sources

	Preceptor		
	1	2	3
Interaction			
P talks, S listens	.99	.90	.53
P questions, S answers	.50	.74	.38
P questions, P answers	.16	.11	.00
S talks, P listens	.57	.58	.51
S questions, P answers	.62	.46	.25
S questions, S answers	.04	.04	.04
Other	.22	.17	.09
Topics			
Instand word recognition	.48	.46	.35
Decoded word recognition	.40	.43	.46
Oral reading	.24	.23	.33
Reading comprehension	.27	.32	.28
Message comprehension	.22	.13	.21
Word comprehension	.27	.33	.21
Attention	.20	.16	.12
Cue collection	.60	.55	.12
Other	.62	.54	.77
Topic Sources			
P personal experiences	.55	.17	.04
S personal experiences	.27	.20	.00
Dx materials in general	.19	.38	.00
Dx Rx tables	.06	.39	.14
Dx Rx checklists	.00	.07	.30
Dx Rx glossary	.00	.00	.05
Dx Rx decision aid	.01	.45	.21
Cases in general	.19	.04	.02
Simulated cases	.01	.43	.46
Mini-cases	.00	.04	.00
Real cases known by S	.60	.20	.14
Real cases known by P	.30	.00	.00
Anecdotes	.48	.12	.02
Tests in general	.18	.30	.02
Standardized tests	.62	.45	.21
Non-standardized tests	.08	.35	.11
Textbooks, printed documents	.10	.18	.07
MORAL	.38	.04	.04
Other	.29	.24	.16

Note: P: preceptor, S: student, Dx: diagnosis, Rx: remediation

observational studies because of the availability of decision aids. Further, it was expected that agreement would increase as a result of training.

The mean agreement across cases for the total diagnosis is shown in Figure 1 and in the Appendix E. Three different sets of individual diagnostic agreement data are represented:

1. agreement among the three preceptors, measured at the end of the training program;
2. agreement of students with their own preceptors; and
3. agreement among students across training groups on a given case.

Student agreements with themselves and their preceptors are shown for pretests, midtests, posttests, and transfer tests. Agreement among preceptors and among students reflected the influence of the decision aid. As can be seen in Figures 1 and 2, preceptor agreement is markedly higher than the mean value obtained in the series of observational studies (Porter = 0.37 vs. 0.08; Phi = 0.46 vs. 0.03). For the untrained students, initial diagnostic

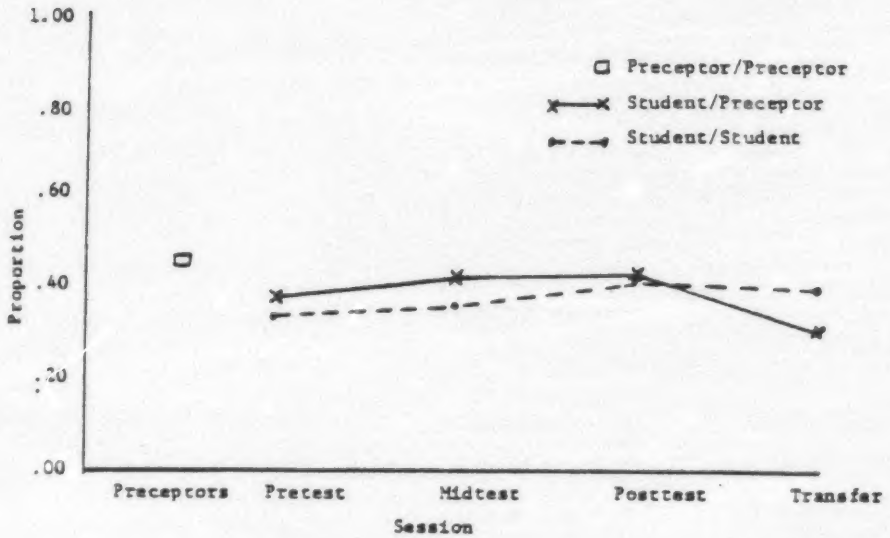

Figure 1. Total diagnosis (mean Phi)

agreement was higher than that obtained by the experienced specialists in the observational studies (Porter = 0.26 vs. 0.08; Phi = 0.34 vs. 0.03). In addition, the students showed modest gains across the pre-, mid-, and posttests both for student/student and student/preceptor agreement. On the pretest, the students agreed more with their preceptors than with one another. After training, this difference decreased to zero on the posttest. On the transfer test, individual agreement of student with student was maintained, but the individual agreement of student with preceptor actually declined to the pretest level.

To summarize, the overall improvement due to training and decision aids is impressive compared to that of clinicians working from their traditional training and without decision aids. The improvement transfers to cases not previously diagnosed and influences practitioners and experienced clinicians. Further, the student/student agreement shows sustained improvement from pretest to transfer test. However, while the student/preceptor agreement shows improvement from pre- to midtest, a puzzling decline appears from midtest, to posttest, to transfer test.

We analyzed the data further to find an explanation for this unexpected decrease in agreement between students and preceptors. First, we wanted to determine if the decrease in agreement held for both critical reading performances and causal factors. The data for critical reading performances are shown in Figure 2 and Table 5. The data reveal that the effect (i.e., the higher level of agreement of students with each other than with their preceptors) not only held for the transfer test but the posttest as well. As with total diagnosis, student/student agreement on the pretest was lower than student/preceptor agreement.

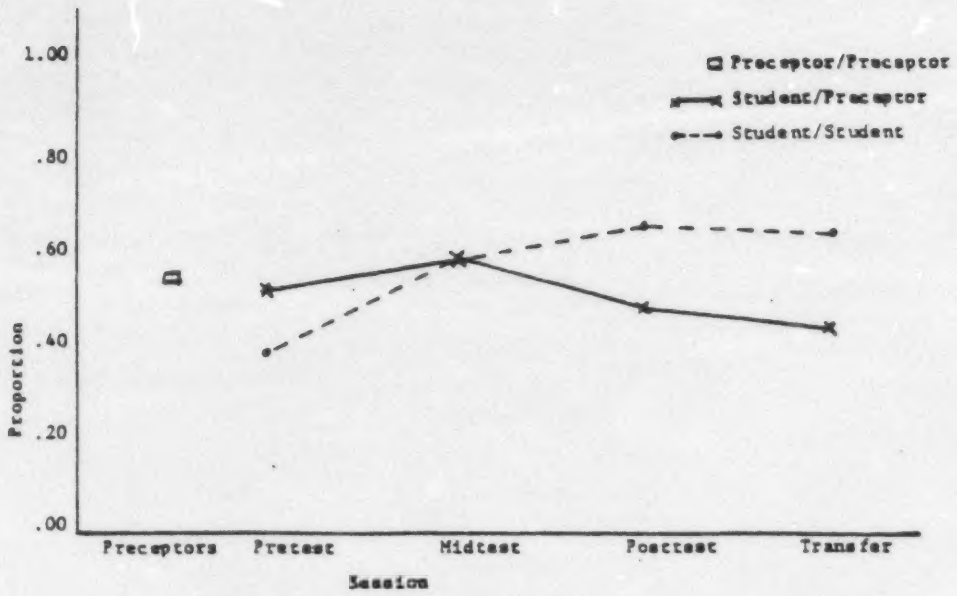

Figure 2. Critical reading performances (mean Phi)

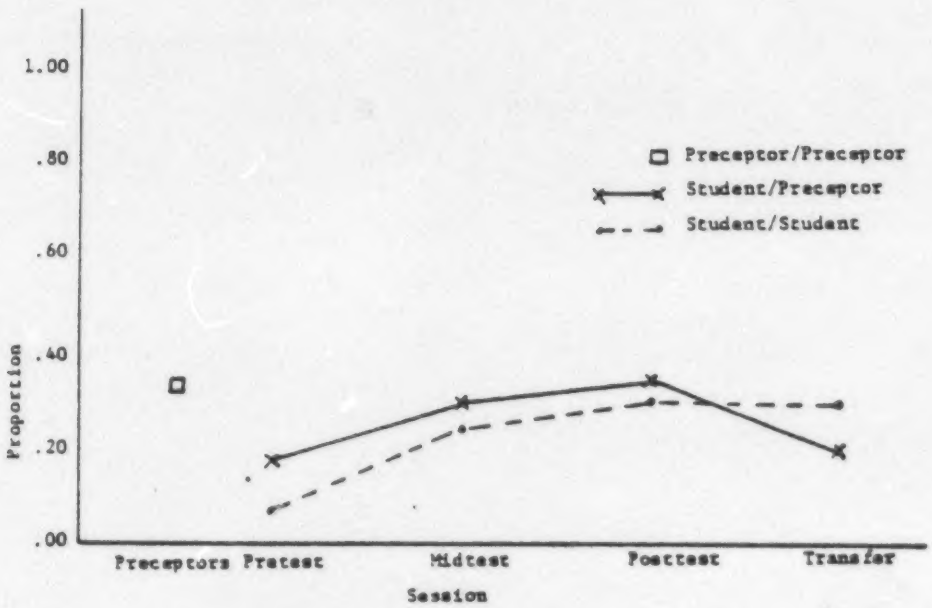

Figure 3. Causal factors (mean Phi)

Table 5

Commonality Results for the

Critical Reading Performances

CRITICAL PERFORMANCE	STEPHEN			DONALD			MIKE			DAN			HARRY		
	P	SP	ST	P	SP	ST	P	SP	ST	P	SP	ST	P	SP	ST
Instant word rec. (S)	.00	.00	.00	.00	.00	.00	.67	.67	.33	.00	.00	.00	.33	.00	.00
Instant word rec. (W)	1.00	1.00	1.00	1.00	1.00	1.00	.33	.33	.67	1.00	1.00	1.00	.67	1.00	1.00
Decoded word rec. (S)	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Decoded word rec. (W)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Meaning vocab. (S)	1.00	.33	.00	.67	.00	.00	1.00	.33	.33	.00	.67	.00	.00	.00	.00
Meaning vocab. (W)	.00	.67	1.00	.33	1.00	1.00	.00	.67	.67	.00	.33	1.00	.67	1.00	1.00
Oral reading (S)	.00	.00	.00	.00	.00	.00	.00	.00	.00	.33	.00	.00	.00	.00	.00
Oral Reading (W)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.67	1.00	1.00	1.00	1.00	1.00
Reading comprehen. (S)	.33	.00	.00	.33	.00	.00	.00	.00	.00	.67	.33	.33	.00	.00	.00
Reading comprehen. (W)	.67	.67	1.00	.67	1.00	1.00	1.00	1.00	1.00	.00	.67	.67	1.00	1.00	1.00
Listening compre. (S)	1.00	1.00	.67	.33	.00	.33	1.00	.00	.00	1.00	1.00	.67	.00	.00	.00
Listening compre. (W)	.00	.00	.33	.67	1.00	.67	.00	1.00	1.00	.00	.00	.33	1.00	1.00	1.00
Attention/Motiv. (S)	.33	.33	.33	.00	.00	.33	.67	.00	.00	.33	.67	.67	1.00	.33	.00
Attention/Motiv. (W)	.67	.67	.67	1.00	1.00	.67	.33	1.00	1.00	.67	.33	.33	.00	.67	1.00

Note. P: preceptor, agreement, SP: student agreement on posttest, ST: student agreement on transfer test. Names of cases are not those of actual children whose profiles were used.

Figure 3 shows the data for causal factors. As can be seen, the profiles parallel those for total diagnosis, except that there was generally lower agreement on causal factors than on total diagnosis.

Several hypotheses were suggested to account for the results showing that student/student agreement consistently outstripped student/preceptor agreement. The first was that students might have fewer diagnostic categories than the preceptors, and thus perhaps had higher agreement because they stuck to the more obvious observations and provided less details. This hypothesis had to be rejected. Students did use fewer categories than the preceptors on the pretest (33 vs. 38), but they actually used substantially more categories on the posttest (45) and transfer test (50).

A second hypothesis about the students' higher reliability on the transfer test was that the students' diagnoses might have been simpler than those of the preceptors (i.e., might have contained more statements about critical reading performances and fewer about complex matters of causality).

To explore this possibility we identified the diagnostic categories used by all clinicians or all students. There were 66 such categories, about half the total number of categories on the checklist. Using this as a base, we identified (1) those categories agreed upon by all preceptors but not by all students; and (2) those agreed upon by all students but not by all preceptors. These categories were identified for pretest, midtest, posttest and transfer test. Finally, the categories were divided into those dealing with critical performances and those dealing with causal factors.

The proportion of diagnostic categories identified as critical performances for the pretest, posttest, and transfer tests indicates the relative use of these categories by students and preceptors.

As hypothesized, the diagnostic categories most agreed upon by students on the pretest were mainly critical reading performances (0.87 for students vs. 0.41 for preceptors). However, on the posttest and transfer tests the diagnostic categories most agreed upon by students included fewer critical reading performances than did the categories most agreed upon by preceptors (0.27 vs. 0.39 on the posttest and 0.39 vs. 0.47 on the transfer test).

A third hypothesis was that the students might have formed cohort groups, discussed the same cases, and in this way increased their agreement with each other and lowered their agreement with their preceptors. However, the agreement statistics were calculated between students from *different* training groups. Since the groups met on different days, had a different case order, and practiced on cases only within their groups, it is highly unlikely that there was *any* cross-group practice to account for the increased agreement exhibited on the post and transfer tests. Thus we are left with the hypothesis that the students actually had become more systematic in diagnosing cases than their preceptors.

Commonality Results

Further evidence of the impact of training and decision aids on the agreement in content of diagnoses is offered by commonality. Commonality is a measure of group agreement. The diagnostic categories with high commonalities are those which best characterize the group diagnosis of a given case. Overall, the mean commonality on the pretest was higher in this study than in the observational studies (0.54 on the total diagnosis vs. 0.20 for the observational studies). Commonality also increased from pre- to posttests (from 0.54 for the total diagnosis to 0.67). These results confirm the contribution of model-based training and decision aids to group agreement. In this, as in

all prior studies, the critical reading performances figured prominently in the group diagnosis (see Table 5).

As may be seen from the table, most of the commonalities are 1.0, meaning that there was complete agreement on the critical reading performances seen as characterizing the case. However, group agreement was not uniformly distributed over the seven critical reading performances. Decoded word recognition and listening comprehension had the highest group agreement; meaning vocabulary and attention/motivation had the lowest. One possible cause of this is that the simulated cases lack hard data on the latter two factors. Another possible cause is that these indicators are inherently ill-defined within the field of reading.

The next analysis concerns the causal factors most frequently mentioned by preceptors and students across all five cases (Table 6).

Analysis of the table suggests that four major types of causal factors are most frequently agreed upon across the cases: (1) interactions of critical reading performances, (e.g., instant word recognition and decoded word recognition as interfering with oral reading proficiency); (2) subskills of the critical reading performances (e.g., sound-symbol associations for vowels and segmentation of syllables as causes for decoded word recognition; (3) overall perceptual problems (e.g., with visual memory and visual discrimination); and (4) general factors that affect learning (e.g., the amount of practice, motivation, etc.).

For the students in this training study, at least, the critical reading performances not only were important as diagnostic categories, but served as the foundation for examining causes of reading problems.

Table 6

Commonality Results for the

Most Frequently Mentioned Causal Factors

CATEGORY	STEPHEN			DONALD			MIKE			DAN			HARRY		
	P	SP	ST	P	SP	ST	P	SP	ST	P	SP	ST	P	SP	ST
116 Practice (W)	.67	.33	.00	1.00	1.00	.67	1.00	.67	.67	.67	.33	.00	.33	1.00	.33
26 Sound/Symbol Assn vowels (W)	.33	1.00	.33	.33	.33	.33	1.00	.00	.33	1.00	.67	.67	.00	.00	.33
114 Learning Factors (W)	.67	.33	.00	.67	1.00	.67	1.00	1.00	.67	1.00	.67	.00	.33	1.00	.33
28 Segmentation of Syllables (W)	.33	1.00	.67	1.00	.67	1.00	.33	1.00	1.00	1.00	.00	1.00	.33	1.00	.33
62 IWR as cause of poor OR (W)	.67	.67	1.00	.67	1.00	1.00	.33	.33	.67	.67	1.00	.67	.33	.67	.67
64 DWR as cause of poor OR (W)	.67	.67	1.00	.67	1.00	1.00	.33	1.00	.67	.67	1.00	.67	.33	.67	.67
12 Visual memory (W)	.33	.67	.67	.67	.33	1.00	.00	.00	.33	.33	1.00	1.00	.00	.67	.33
10 Vis. Discrim. (W)	.67	.33	.67	.67	1.00	.00	.00	.00	.33	1.00	.33	.33	.00	.00	.33
4 Dolch List (W)	1.00	1.00	.67	.33	.33	.00	.00	.33	.33	.67	.67	.33	.00	.67	.00
56 Meaning Vocab--- definitions Inad (W)	.00	.00	.67	.33	.67	.33	.00	.00	.67	.00	.33	.33	1.00	.33	1.00

Note. P: preceptor, SP: student agreement on posttest, ST: student agreement on transfer test. Names of cases are not those of actual children whose profiles were used.

Summary and Discussion

An earlier series of observational studies (Vinsonhaler et al., 1983) showed very low individual diagnostic agreement of experienced clinicians with each other and with themselves on the same simulated case of reading difficulty. However, the results of three training studies in reading diagnosis showed that diagnostic reliability (agreement) can be raised from approximately zero to about 0.66 through improved training. The training included (1) instruction on a model of the reading process, (2) decision aids based on the model of process, and (3) practice with feedback on simulated cases presented on a minicomputer (DEC PDP8).

The first study examined training not based on an explicit model of the reading process. Instead, training focused on (1) decision aids to make cue collection and diagnostic reporting a routine and systematic process and (2) practice with decision aids on simulated cases. Instruction took place within a diagnostic course for graduate students in reading. Some of the students were practicing teachers. The results showed a marked increase in agreement of student diagnoses with critical diagnoses compiled by a group of senior reading clinicians.

In the second study, training was based on an explicit model of the reading process emphasizing four critical reading performances. The content of instruction included the model of process and applications of the model to diagnosis. The decision aids (including a diagnostic record form and checklist) were explicitly based on the model of process. The practice with simulated cases on the minicomputer included instructor feedback based on the model. Small-group instruction took place within a reading diagnosis course for graduate students, all of whom were teachers with prior graduate training in reading. Results were analyzed separately for agreement on critical

reading performances and on causal factors. Individual diagnostic agreement on the pretest was notably higher in this study than in the series of observational studies (e.g., a mean Phi of .26 vs. .03 in the earlier studies). This pretest difference can probably be attributed to the use of decision aids since all other testing conditions were identical. Further, students' individual diagnostic agreement improved from pre- to posttest on both checklists (e.g., from 0.26 to 0.38 on the data-base checklist) indicating further improvement as a result of model-based training and practice.

The final study examined the impact on individual diagnostic agreement of more refined, extended, and better controlled model-based training than in the second study. New instructional materials and new diagnostic record forms and checklists were developed from a model that had seven critical reading performances. Practice was scheduled on the minicomputer and feedback provided by preceptors for students' diagnoses of five, rather than four, cases. Small-group instruction in diagnosis was provided to practicing teachers with no prior graduate training in reading. Results were analyzed separately comparing the diagnosis of student with student, and student with preceptor on (1) total diagnosis, (2) critical reading performances, and (3) causal factors. The results of the third study confirmed the findings of the second study except that agreements were generally much higher (e.g., pretest student/student agreement on the critical reading performances was 0.39 and posttest agreement was 0.66).

We see three major implications of this work. First, diagnosis in reading can have all the virtues proposed for it in the literature provided its reliability and validity can be established. Second, for reliability to be improved, present methods of diagnostic training must be modified to include the type of training reported here. Third, if the validities of diagnoses are

to be established, empirical studies of remediation based on diagnoses of known reliability must be performed. The authors are presently conducting such validity studies. Methodological issues are discussed by Wagner (1982), and results will be reported in subsequent papers.

Recommendations

Based on our research, we would propose the following method for implementing model-based training in existing inservice and preservice programs. First, select a model of the reading process that lends itself to directing specific diagnostic and remedial actions. The skills-based model chosen here is but one example. Second, create (1) instructional materials that teach the model directly and (2) decision aids that help the student apply the model to diagnostic decision making. Those used in these studies provide examples of such materials and aids.

Third, provide the means to (1) give practice on simulated cases with individualized, model-based feedback; and (2) monitor changes in reliability for pre- and posttesting and possibly for certification testing.

The computer-based method used in the present study worked well. Sets of programs for presenting simulated cases on small computers have proven both time and cost effective. Versions of these programs are under development for low-cost microcomputers (e.g., in BASIC for the Apple II Plus).

Finally, all these resources can be integrated easily into existing courses in reading diagnosis or via an additional clinical practicum.

In summary, our earlier studies uncovered severe problems with diagnostic reliability in reading. The studies reported here have documented a potential solution to the reliability problem based on changes in training. Responsibility for the long-range solution to the problem rests with the educational community.

References

Bateman, B. (1971). The role of individual diagnosis in remedial planning reading disorders. In Calkins (Ed.), Reading Forum (NINDS Monograph No. 11). Bethesda, MD.

Bordage, G. (1982). The cognitive representation of medical knowledge: Categories and prototypes. Unpublished doctoral dissertation, College of Human Medicine, Michigan State University.

Carter, H.L., & McGinnis, D.J. (1970). Diagnosis and treatment of the disabled reader. Toronto: MacMillan.

Cochrane, A.L., & Garland, L.H. (1952). Observer error in the interpretation of chest films: International investigation. Lancet, 2, 505-509.

Collen, M., Rubin, L., Neyman, J., Dantzig, G., Baer, R., & Siegelaub, A. (1964). Automated multiphasic screening and diagnosis. American Journal of Public Health, 54(5).

Copp, H. (1976). Physiology in the curriculum. In E.F. Purcell (Ed.), Recent trends in medical education. New York: Macy Foundation, 151-155.

Cureton, D., Stewart, G., & Patriarca, L. (1980). Diagnosis and remediation in reading. Unpublished paper, Institute for Research on Teaching, Michigan State University.

DeDombal, F.T., Leaper, D., Horrocks, J., Staniland, J., & McCann, A. (1974). Human and computer-aided diagnosis of abdominal pain: Further report with emphasis on performance of clinicians. British Medical Journal, 1, 376-380.

DeGowin, E.L., & DeGowin, R.L. (1976). Bedside diagnostic examination (3rd ed.). New York: MacMillan.

Ekwall, E. (1976). Diagnosis and remediation of the disabled reader. Boston, MA: Allyn & Bacon.

Elstein, A., Shulman, L.S., & Sprafka, S. (1978). Medical problem solving: An analysis of clinical reasoning. Cambridge, MA: Harvard University Press.

Fletcher, C.M. (1952). Clinical diagnosis of pulmonary emphysema: Experimental study. Proceedings of the Royal Society of Medicine, 45, 577-584.

Garland, L.H. (1959). Studies on the accuracy of diagnostic procedures. American Journal of Roentgenology, 82, 25-38.

Gil, D., Polin, R., Vinsonhaler, J., & VanRoekel, J. (1980). The impact of training on diagnostic consistency (Research Series No. 67). East Lansing, MI: Institute for Research on Teaching, Michigan State University.

Gil, D., Wagner, C.C., & Vinsonhaler, J.F. (1978). Simulating the problem solving of reading clinicians (Research Series No. 30). East Lansing, MI: Institute for Research on Teaching, Michigan State University.

Guszak, F.J. (1972). Diagnostic reading instruction in the elementary school. New York: Harper & Row.

Harris, A.J. (1977). Ten years of progress in remedial reading. Journal of Reading, 21(1), 29-35.

Harris, A.J. (1972). The diagnosis of reading disabilities. In A.J. Harris & E.R. Sipay (Eds.), Readings on reading instruction (2nd ed.). New York: David McKay.

Johnson, D. (1976). The medical student, 1975. In E.F. Purcell (Ed.), Recent trends in medical education. New York: Macy Foundation, 37-54.

King, D. (1976). Pathology in the curriculum. In E.F. Purcell (Ed.), Recent trends in medical education. New York: Macy Foundation, 156-164.

Lee, A., & Weinshank, A. (1978). Case production and analysis: CLIPIR pilot observational study of reading diagnosticians (Research Series No. 14). East Lansing, MI: Institute for Research on Teaching, Michigan State University.

Lerner, J., & Schuyler, J. (1973, August). Computer applications in the field of learning disabilities (Final Report), U.S. Dept. of Health, Education & Welfare.

Monroe, M. (1968). General principles of diagnosis of reading disabilities. In D.G. Schubert & T.L. Targerson (Eds.), Readings in reading: Practice theory research. New York: Thomas Y. Crowell Company.

Natchez, G. (Ed.) (1978). Children with reading problems. New York: Basic.

Norman, G.R., & Tugwell, P. (1981). The validity of simulated patients. Presentation at the 1981 Research in Medical Education Conference, Washington, D.C.

Otto, W., McMenemy, R.A., & Smith, R.J. (1973). Corrective and remedial teaching (2nd ed.). Boston, MA: Houghton Mifflin.

Paton, B.C. (1957). The accuracy of diagnosis of myocardial infarction. American Journal of Medicine, 23, 761-768.

Polin, R.M. (1981). A study of preceptor training of classroom teachers in reading diagnosis (Research Series No. 110). East Lansing, MI: Institute for Research on Teaching, Michigan State University.

Prior, J.A., Silberstein, J.S., & Stang, J.M. (1981). Physical diagnosis: The history and examination of the patient (6th ed.). St. Louis, MO: C.V. Mosby.

Puck, T. (1976). Cell biology in the curriculum. In E.F. Purcell (Ed.), Recent trends in medical education. New York: Macy Foundation, 144-150.

Rabinovitch, R. (1965). Differential diagnosis in children with reading retardation. In J.H. Root (Ed.), Diagnostic teaching: Methods and materials. Syracuse, NY: Syracuse University, School of Education.

Roos, T. (1976). Teaching the biological sciences to premedical students. In E.F. Purcell (Ed.), Recent trends in medical education. New York: Macy Foundation, 8-20.

Shapiro, E., & Lowenstein, L. (Eds.). (1979). Becoming a physician: development of values and attitudes in medicine. Cambridge, MA: Ballinger.

Simpson, M. (1972). Medical education: A critical approach (Chapter 7). London: Butterworth.

Smith, C.B. (1969). Correcting reading problems in the classroom. Newark, DE: International Reading Association.

Smith, C.B., Carter, B., & Dapper, G. (1970). Treating reading difficulties: The role of the principal, teacher, specialist, administrator. Washington, D.C.: U.S. Department of Health, Education and Welfare.

Spache, G.D. (1969). Integrating diagnosis with remediation in reading (Chap. 2). In H. Newman (Ed.), Reading disabilities. New York: Odyssey.

Spache, G.D. (1976). Diagnosing and correcting reading disabilities. Boston, MA: Allyn & Bacon.

Spache, G.D., & Spache, E.B. (1973). Reading in the elementary school (3rd ed.). Boston, MA: Allyn & Bacon.

Strang, R. (1964). Diagnostic teaching of reading. New York: McGraw-Hill.

Vinsonhaler, J.F., Weinshank, A.B., Wagner, C.C., & Polin, R.M. (1983). Diagnosing children with educational problems: Characteristics of reading and learning disabilities specialists, and classroom teachers. Reading Research Quarterly, 18(2), 134-164. (Also available as Research Series No. 117, East Lansing, MI: Institute for Research on Teaching, Michigan State University, 1982.)

Wagner, C.C. (1982). Learning from experience: Applying the clinical and epidemiological research paradigms to the study of diagnosis and treatment in reading. Unpublished doctoral dissertation, Michigan State University.

Weed, L. (1976). A new paradigm for medical education. In E.F. Purcell (Ed.), Recent trends in medical education. New York: Macy Foundation, 55-93.

Weinshank, A.B. (1982). The reliability of diagnostic and remedial decisions of reading specialists. Journal of Reading Behavior, 14(1), 33-50.

Weinshank, A.B., Cureton, D., & Blatt, G. (1980). A model of reading and learning to read. Unpublished course material, East Lansing, MI: Institute for Research on Teaching, Michigan State University.

Yerushalmy, J. (1955). Reliability of chest radiography in the diagnosis of pulmonary lesions. American Journal of Medicine, 89, 231-240.

Yerushalmy, J. (1969). The statistical assessment of the variability in observer perception and description of roentgenographic pulmonary shadows. Radiologic Clinics of North America, 7, 381-391.

Appendix A
Cue Inventory for Case 4, Dan

Physical Information

Vision test
Audiometric record

Background Information

School record
Teacher form
School information
Parent form

Assessment Information

Basic sight vocabulary (Dolch)
Sentence completion
Gates-McKillop reading
diagnostic tests
  Recognizing & blending
    commond word parts
  Auditory blending
  Phonic spelling of words
  Giving letter sounds
Auditory discrimination (Wepman)
Durrell list-read series
  Intermediate level vocabulary
  Paragraphs

Assessment Information (Cont.)

Durrell diagnostic analysis of
reading difficulty
  Oral Reading
  Silent Reading
  List. Comprehension
  Word Recognition & Word analysis
  Hearing sounds in words--
    primary
  Visual memory of words--
    primary
  Intermediate Spelling--List 1
  Phonic spelling of words
Achievement test (Iowa Test of
Basic Skills)
  Vocabulary
  Reading
Graded word list (Slosson Oral
  Reading Test)
Reading achievement (Gates-
  MacGinitie)
  Speed/Accuracy
Cognitive ability (Wechsler
  Intelligence Scale for Children)
  Verbal
  Performance
  Full scale

Appendix B

Calculation of Phi Correlation
and Porter Statistic

Clinician i SIMCASE Q, Form One

Clinician i SIMCASE Q, Form Two	PRESENT (+)	ABSENT (-)
	Frequency count of statements present in the domain in both sessions for form one and form two of SIM-CASE A	Frequency count of statements present in the domain present in SIMCASE form two but not in SIMCASE form one B
	Frequency count of statements in the domain present in the session for SIM-CASE form one but not SIMCASE form two C	Absent in both sessions for form one and form two of SIMCASE D

	+	-	
+	a (++)	b (+-)	a + b
-	c (-+)	d (--)	c + d
	a + c	b + d	N

$$\text{Phi} = \frac{(a \times d - b \times c)}{(a + C) \times (b + d) \times (c + d) \times (a + b)}$$

The presence of a large percentage of statements (more than 85%) in the "D" cell (the statement is absent in both sessions) artifically inflated the intercorrelations since it represented, in effect, agreeing to disagree. A statistic development by Professor A. Porter (Institute for Research on Teaching, Michigan State University) was designed to correct for this occurrence, by including in the computation only the values in the A, B, and C cells $\frac{A}{A + B + C}$.

Appendix C
A Portion of a Diagnostic Decision Aid (1980 Study)

Case Name: Dan (Grade 4)

Does the student nave a problem with INSTANT WORD RECOGNITION?

(Circle One) Yes No

On what basis was this decision made?

SORT Score: 2.1
Durrell Word Analysis and Word Recognition = low first grade

If no, then continue with the next problem area on page 3.

If yes, describe the important factors that have contributed to this problem. For each factor, suggest remedial procedures for its improvement. Continue on the next page if required.

1. Describe one factor contributing to the problem with Instant Word Recognition.

Dan has poor visual memory of words.

Suggest remedial procedures for alleviating this factor.

He needs to look at the whole word not just the beginning letters.

2. Describe another factor contributing to the problem with Instant Word Recognition.

Dan does not do enough reading outside of class.

Suggest remedial procedures for alleviating this factor.

Parents need to devise a plan to encourage Dan to read more, possibly using a reward system for the amount of reading he does.

Appendix D
A Portion of the Diagnostic Checklist

Case Name_____
Your Name_____
Date _____

1____Instant Word Recognition Adequate
2____Instant Word Recognition Inadequate
3_____ Basic Sight Words Adequate
4_____ Basic Sight Words Inadequate
5_____ Sight words Learned Via Decoding Adequate
6_____ Sight Words Learned Via Decoding Inadequate
7_____ Experiential Sight Words Adequate
8_____ Experiential Sight Words Inadequate
9_____ Visual Discrimination Adequate
10_____ Visual Discrimination Inadequate
11_____ Visual Memory Adequate
12_____ Visual Memory Inadequate
13_____ Print-Meaning Association Adequate
14_____ Print-Meaning Association Inadequate
15_____ Print-Sound Association Adequate
16_____ Print-Sound Association Inadequate
17_____ Other Adequate
18_____ Other Inadequate
19____Decoded Word Recognition Adequate
20____Decoded Word Recognition Inadequate
21_____ Sound-Symbol Association - Consonants Adequate
22_____ Sound-Symbol Association - Consonants Inadequate
23_____ Sound-Symbol Association - Blends/Diagraphs Adequate
24_____ Sound-Symbol Association - Blends/Diagraphs Inadequate
25_____ Sound-Symbol Association - Vowels/Vowel Patterns Adequate
26_____ Sound-Symbol Association - Vowels/Vowel Patterns Inadequate
27_____ Visual Segmentation into Syllables Adequate
28_____ Visual Segmentation into Syllables Inadequate
29_____ Auditory Segmentation into Syllables Adequate
30_____ Auditory Segmentation into Syllables Inadequate
31_____ Blending of Sounds Adequate
32_____ Blending of Sounds Inadequate
33_____ Adjustment of Blended Sounds to Language Adequate
34_____ Adjustment of Blended Sounds to Language Inadequate
35_____ Use of Root Word Adequate
36_____ Use of Root Word Inadequate
37_____ Use of Prefixes Adequate
38_____ Use of Prefixes Inadequate
39_____ Use of Suffixes Adequate
40_____ Use of Suffixes Inadequate
41_____ Auditory Memory Adequate
42_____ Auditory Memory Inadequate
43_____ Auditory Discrimination Adequate
44_____ Auditory Discrimination Inadequate
45_____ Visual Memory Adequate
46_____ Visual Memory Inadequate
47_____ Visual Discrimination Adequate
48_____ Visual Discrimination Inadequate
49_____ Other Adequate
50_____ Other Inadequate

Appendix E

Mean Student/Student (S/S), Student/Perceptor (S/P), and Perceptor/Perceptor (P/P) Agreement on the Diagnoses of Four Cases.

	Pretest						Midtest						Posttest						Transfer					
	Phi			Porter			Phi			Porter			Phi			Porter			Phi			Porter		
	S/S	S/P	P/P	S/S	S/P	P/P	S/S	S/P	P/P	S/S	S/P	P/P	S/S	S/P	P/P	S/S	S/P	P/P	S/S	S/P	P/P	S/S	S/P	P/P
Total diagnoses	.34	.38	--	.26	.29	--	.36	.43	--	.31	.35	--	.42	.44	.46	.37	.36	.37	.41	.31	--	.35	.28	--
Critical reading performances	.39	.52	--	.55	.63	--	.58	.58	--	.68	.68	--	.66	.48	.54	.73	.61	.64	.64	.44	--	.71	.58	--
Causal factors	.08	.18	--	.08	.13	--	.25	.30	--	.21	.22	--	.31	.35	.34	.26	.26	.24	.38	.20	--	.25	.17	--

Note: Where P/P appears blank, values for posttest P/P agreement were used.